An Institutional Perspective to Rescue Scholarly Orphans

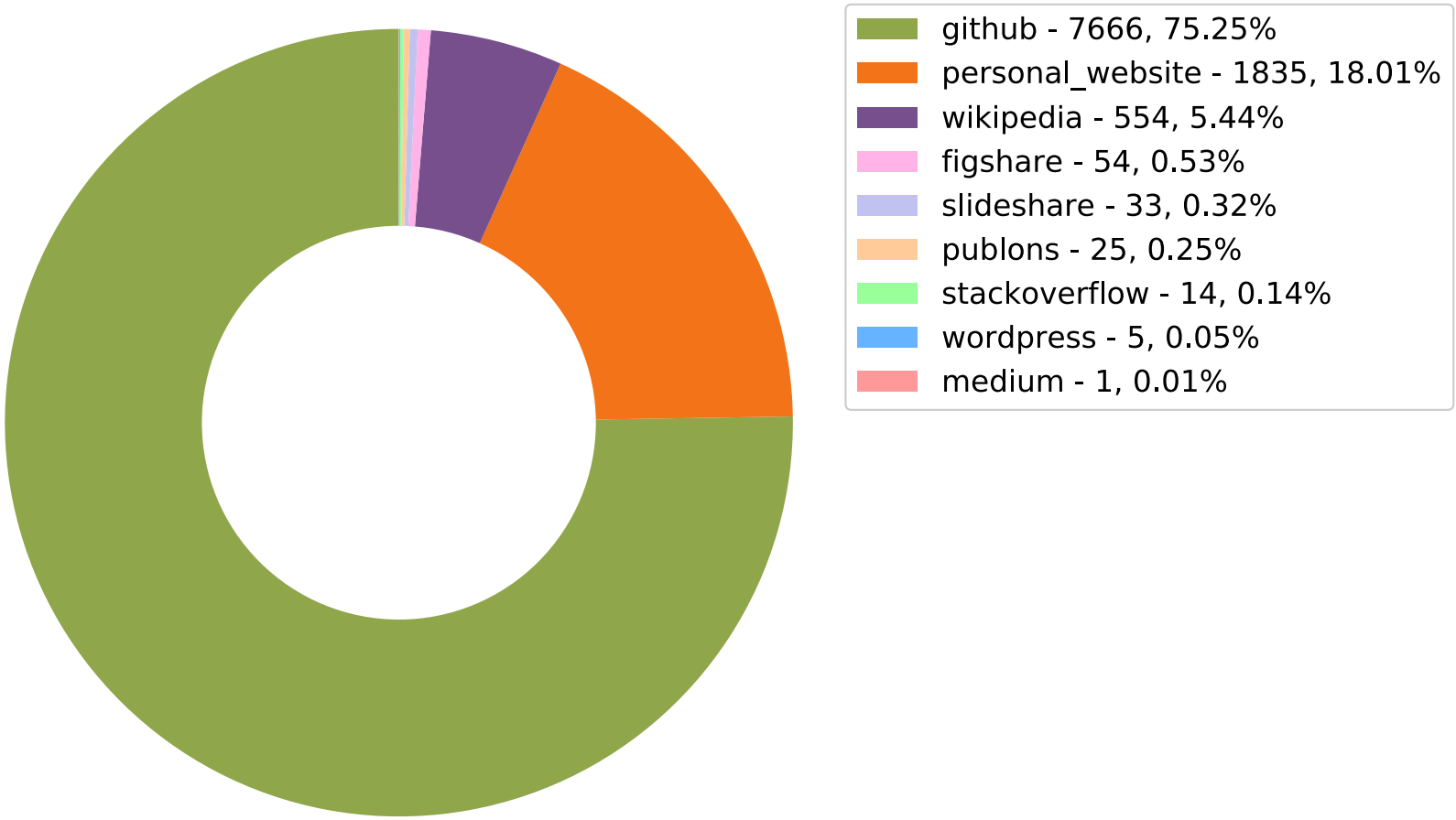
Martin Klein
Los Alamos National Laboratory
@mart1nkle1n
https://orcid.org/0000-0003-0130-2097

The Scholarly Orphans project
is funded by the Andrew W. Mellon Foundation

CC BY SA

Los Alamos NATIONAL LABORATORY

Knowing this would be great!


github - 7666, 75.25%
personal_website - 1835, 18.01%
wikipedia - 554, 5.44%
figshare - 54, 0.53%
slideshare - 33, 0.32%
publons - 25, 0.25%
stackoverflow - 14, 0.14%
wordpress - 5, 0.05%
medium - 1, 0.01%

CC BY SA

Los Alamos NATIONAL LABORATORY

Research and Research Communication on the Web

- Researchers are increasingly using a variety of web platforms for collaboration and communication
- Researchers deposit artifacts in web platforms
- The researchers' institutions are in the dark
- Uncertainty regarding long-term access of scholarly artifacts
- Scholarly artifacts are not systematically archived

CC BY SA

Los Alamos NATIONAL LABORATORY

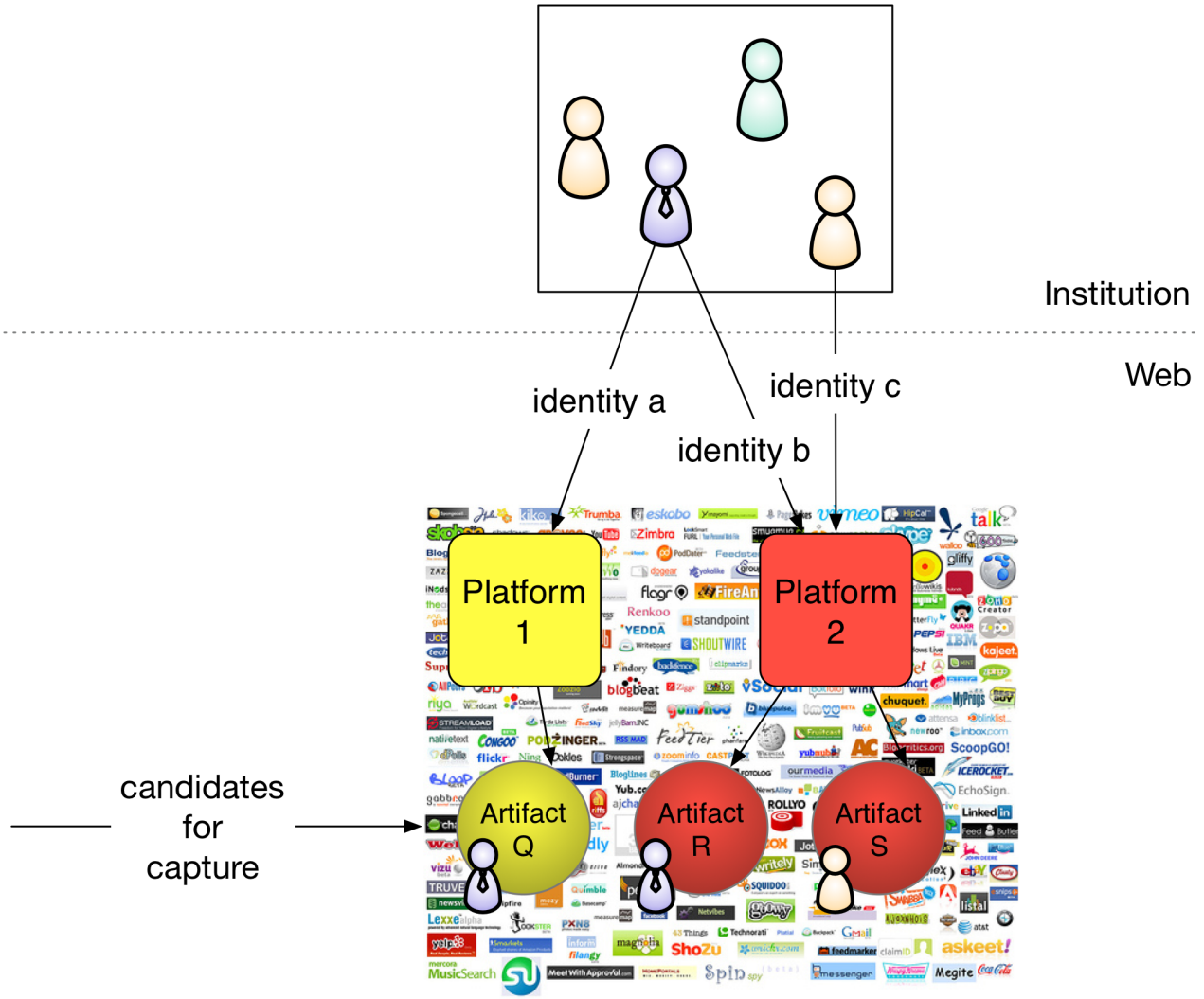
Scholarly Orphans Project

How to faithfully capture Scholarly Orphans for long-term archiving?

CC BY SA

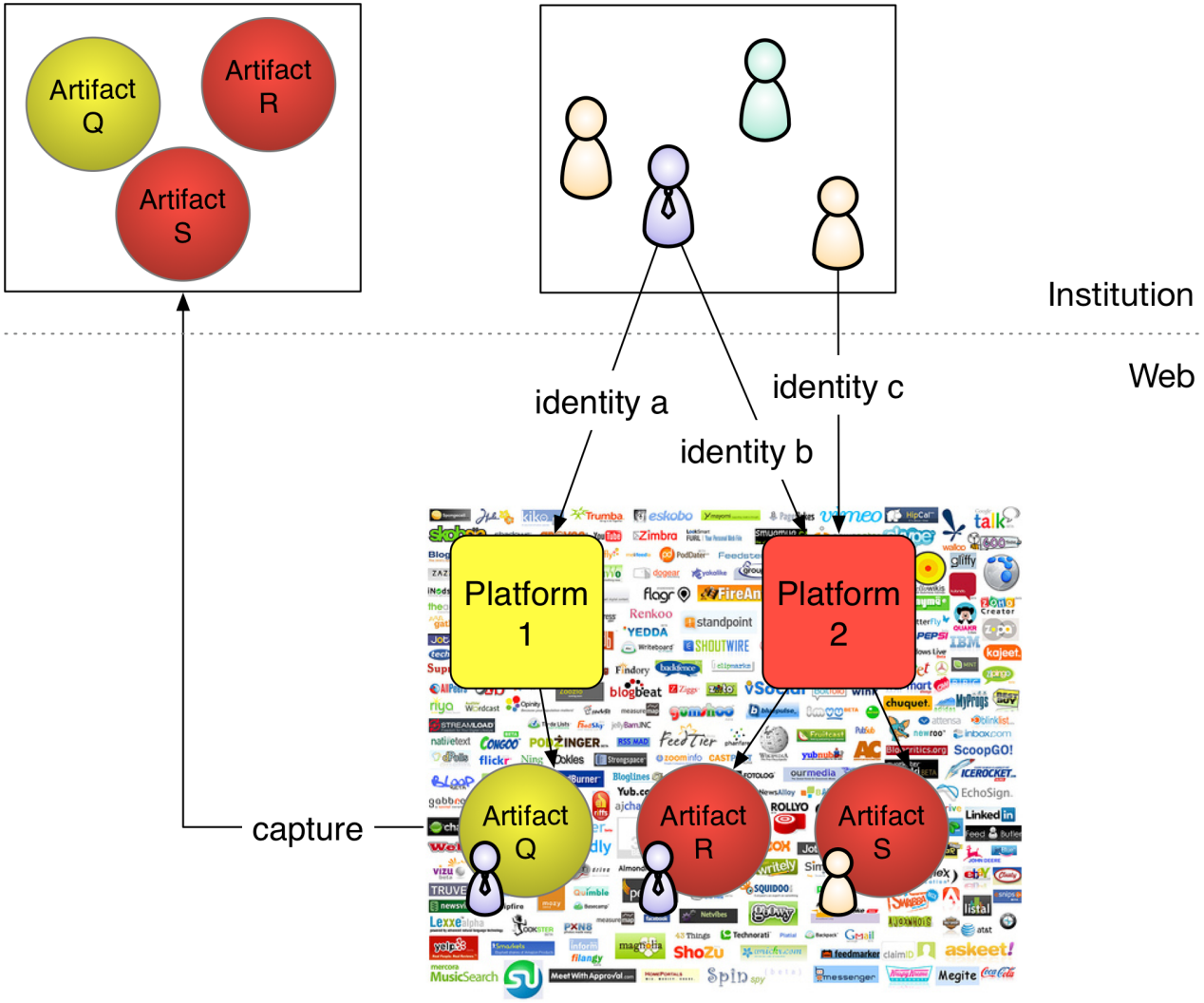
Los Alamos NATIONAL LABORATORY

An Institutional Perspective



CC BY SA
Los Alamos NATIONAL LABORATORY

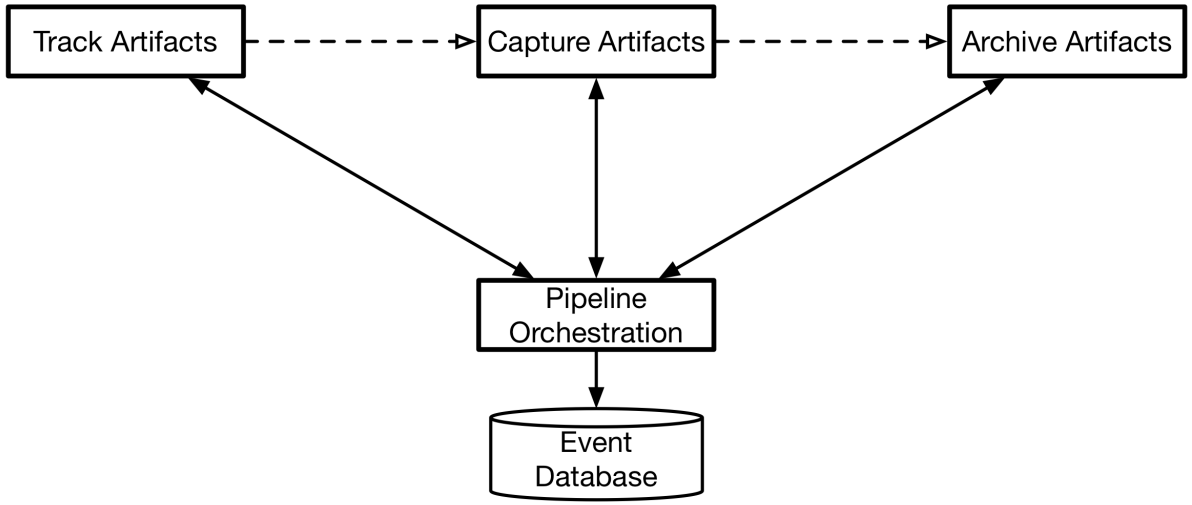
A Web Archiving Perspective



CC BY SA

Los Alamos NATIONAL LABORATORY

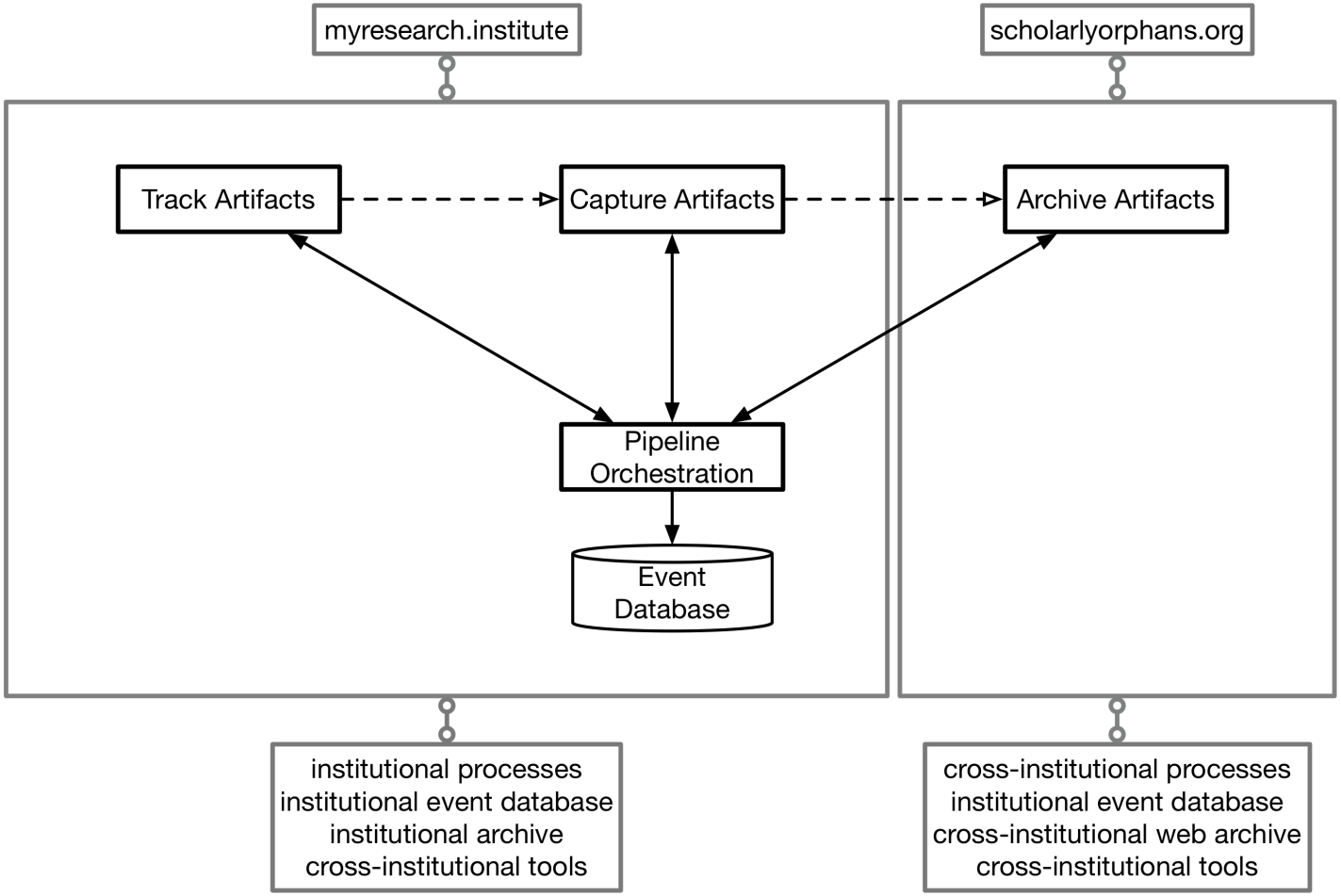
Prototype Pipeline



CC BY SA

Los Alamos NATIONAL LABORATORY

Pilot - myresearch.institute



CC BY SA

Los Alamos NATIONAL LABORATORY

myresearch.institute - Researchers

- Uniquely identified by ORCIDs
- Web identities in multiple portals
- Create various types of artifacts

CC BY SA

Los Alamos NATIONAL LABORATORY

myresearch.institute - Portals

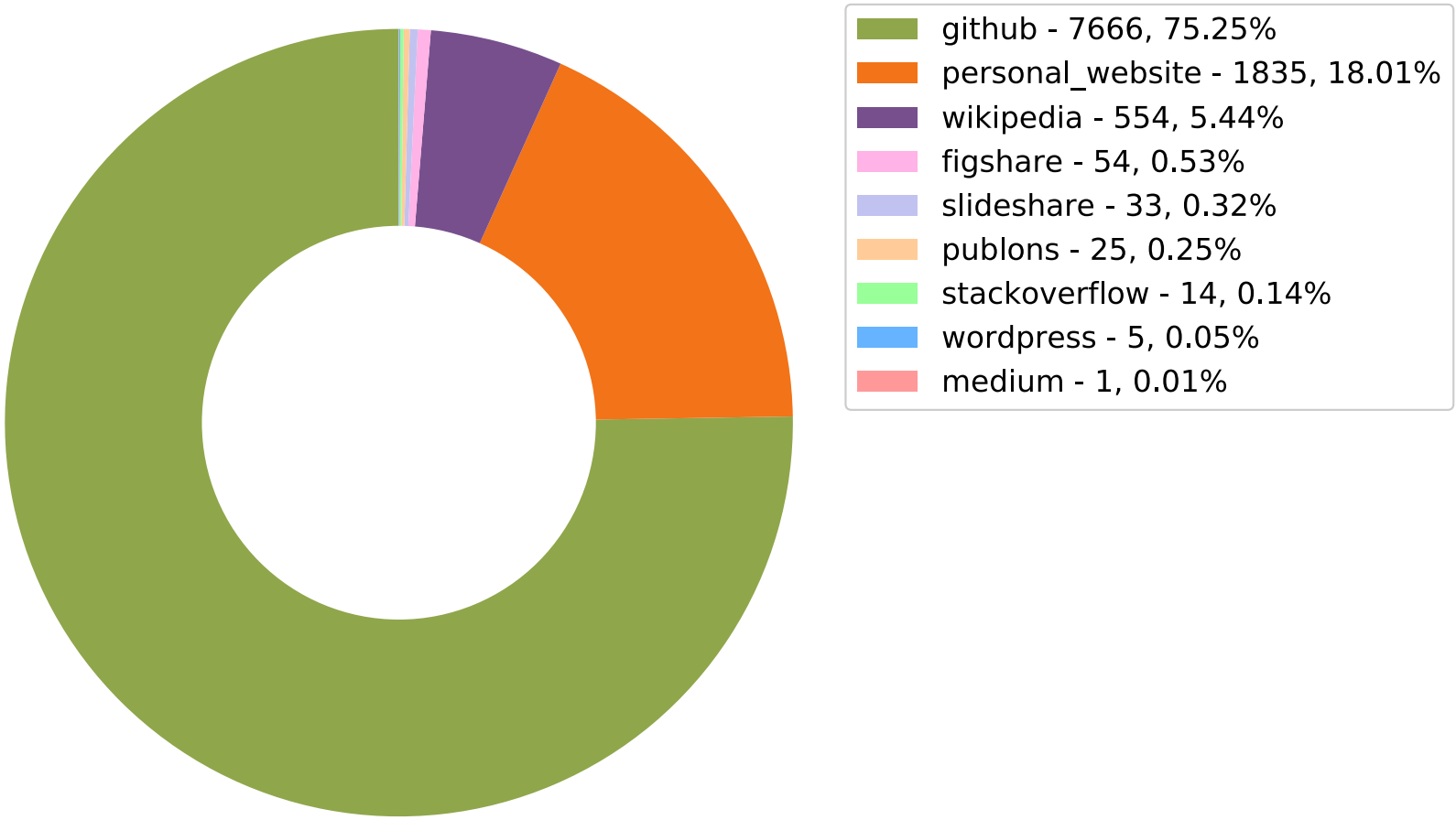
- Tracking started August 27 2018
- Tracking artifacts created starting August 1 2018
- ~17,000 artifacts tracked to date for all 16 researchers

CC BY SA

Los Alamos NATIONAL LABORATORY

myresearch.institute – Statistics

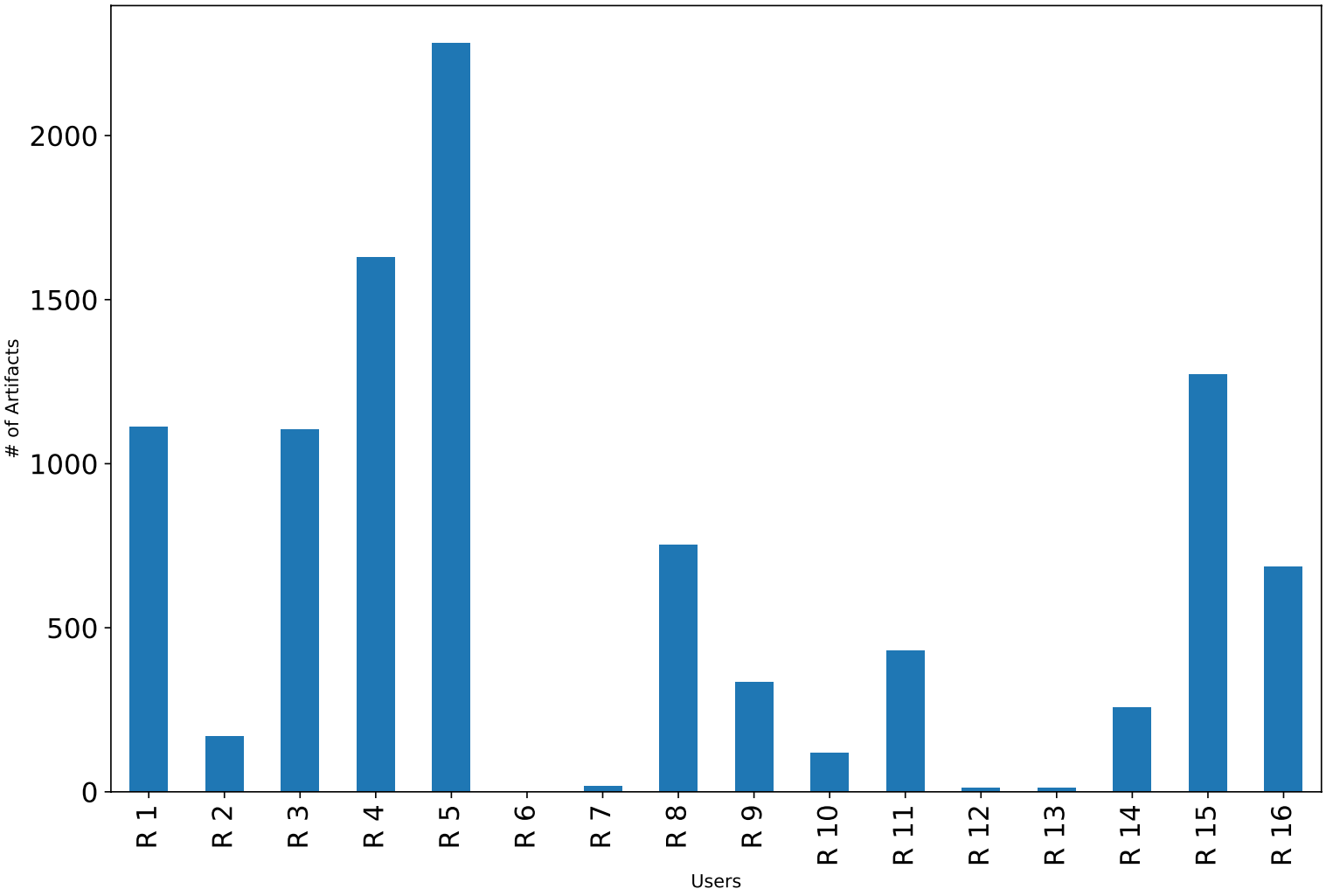
Scholars' Productivity Portal Distribution


github - 7666, 75.25%
personal_website - 1835, 18.01%
wikipedia - 554, 5.44%
figshare - 54, 0.53%
slideshare - 33, 0.32%
publons - 25, 0.25%
stackoverflow - 14, 0.14%
wordpress - 5, 0.05%
medium - 1, 0.01%

CC BY SA

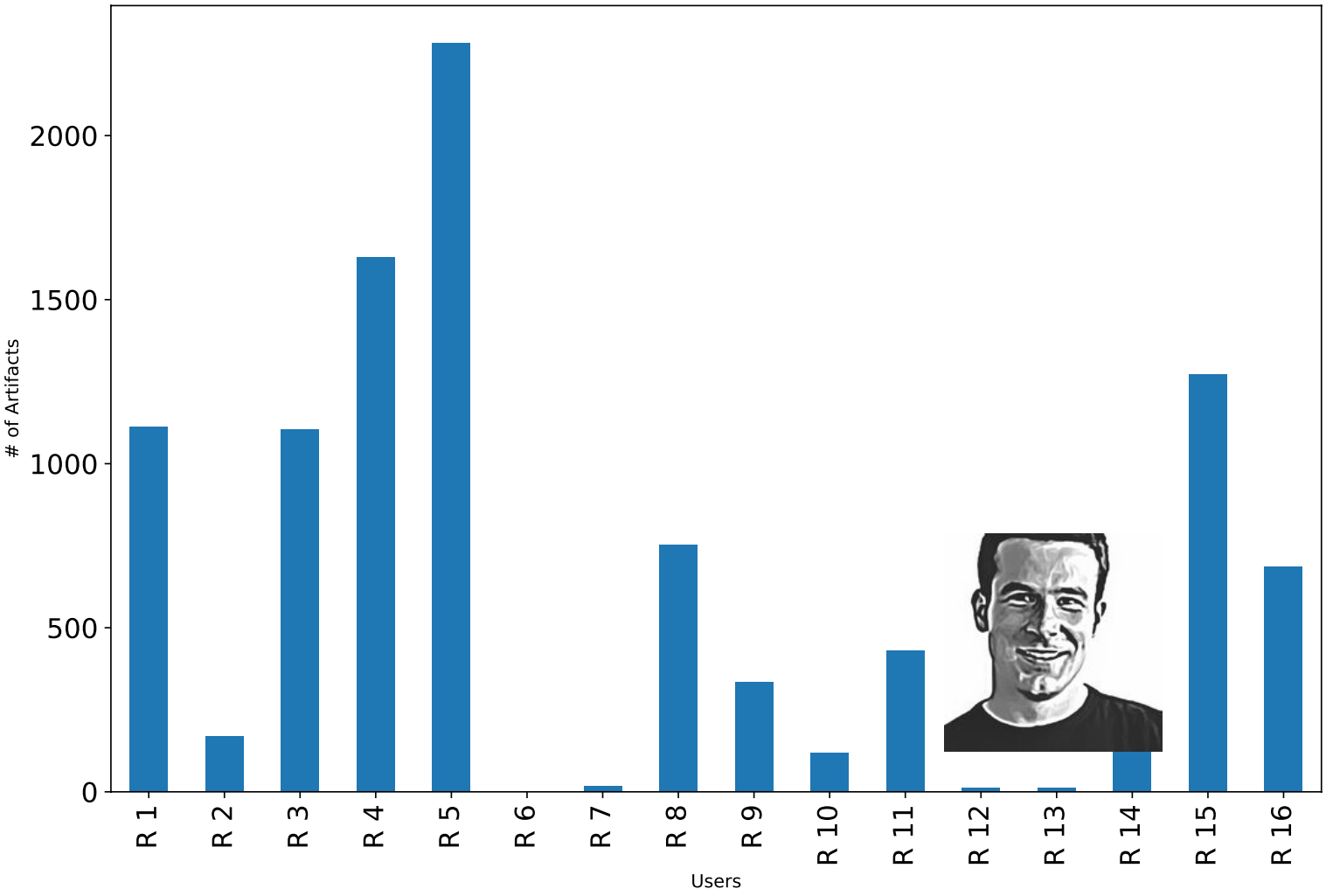
Los Alamos NATIONAL LABORATORY

Researcher Contributions



CC BY SA

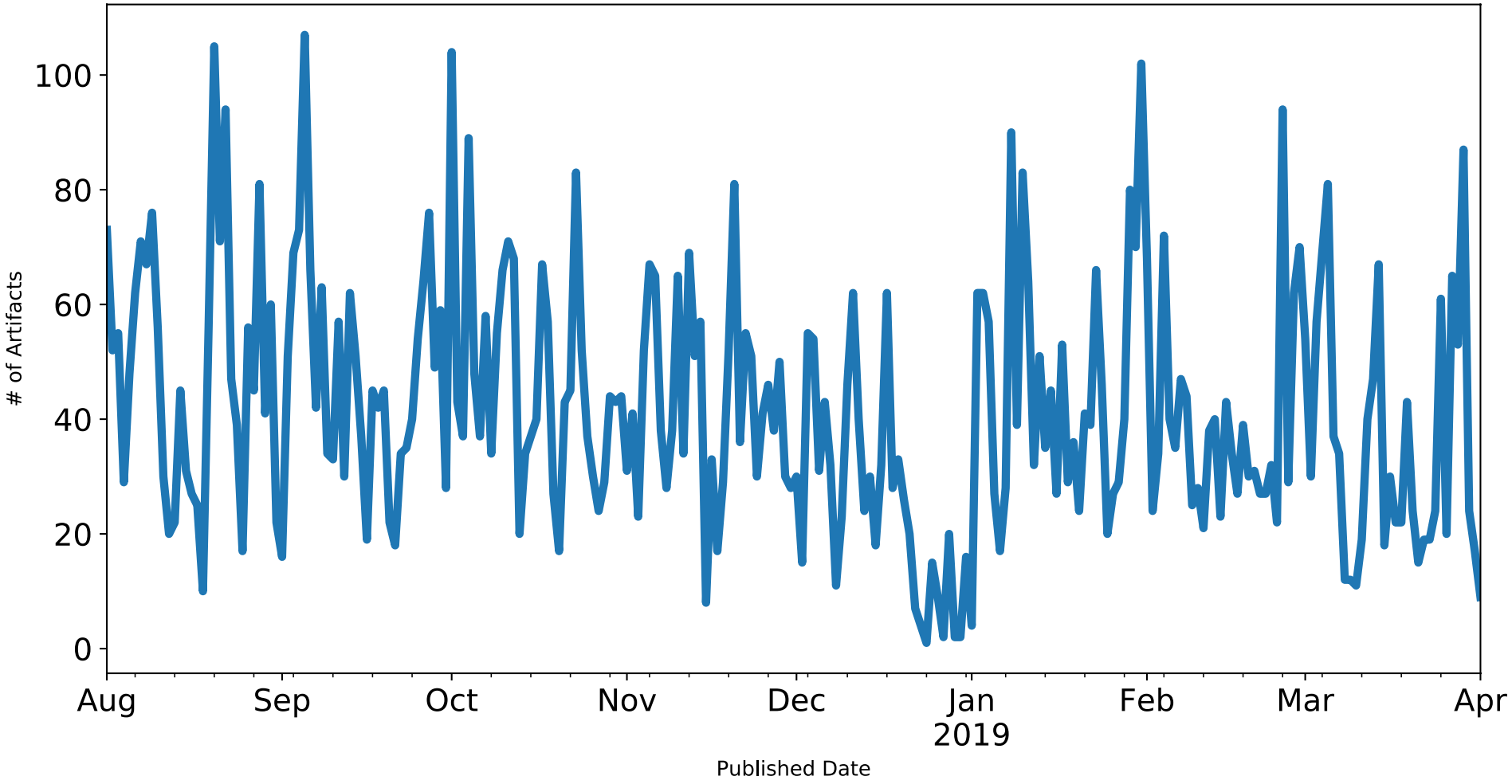
Researcher Contributions



CC BY SA

Los Alamos NATIONAL LABORATORY

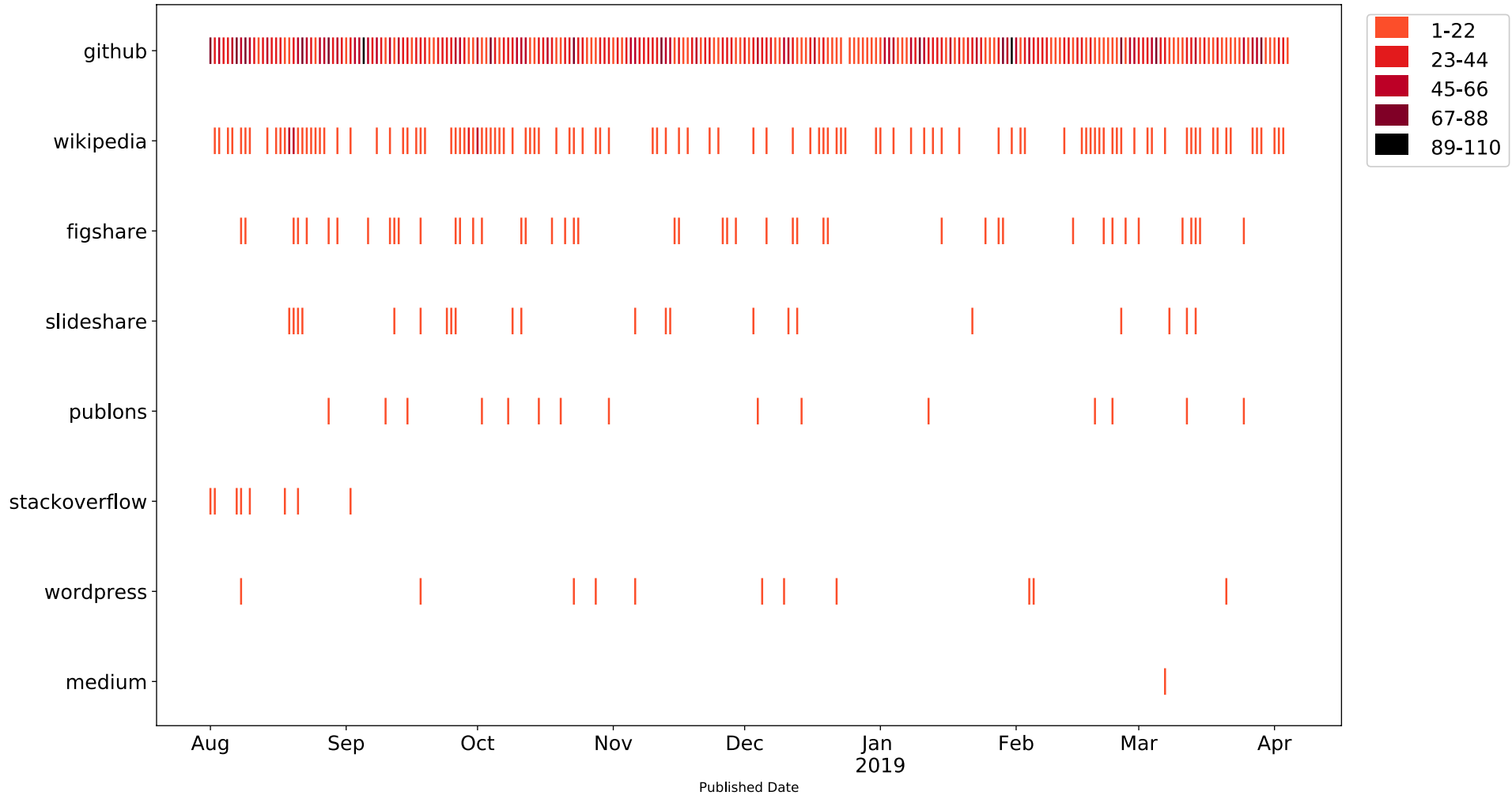
Artifact Frequency



CC BY SA

Los Alamos NATIONAL LABORATORY

Artifact Frequency per Portal



CC BY SA

Los Alamos NATIONAL LABORATORY

Summary

- The Scholarly Orphans project explores an institution-driven approach to capture scholarly artifacts deposited in web portals
    - Artifacts out of scope of existing archival approaches such as LOCKSS, Portico, web archives
    - Institutions have a long shelf life, should be interested in collecting these artifacts, and have feasible scale for identity/ artifact discovery
- Prototype at myresearch.institute illustrates feasibility, opportunities, and challenges of this institutional perspective

CC BY SA

Los Alamos NATIONAL LABORATORY

What Our Researchers Say…

“Ha, this is awesome! Thanks for letting me know - carry on as usual, and feel free to monitor away. I'll try not to change my behaviour or anything now with this new knowledge :)”

“This is fine, since everything you are capturing is public to start with. I also wonder if you know about Software Heritage?”

“I’m very comfortable with being part of this (very important) research project”

“I'm cool with it :-)”

“Interesting project! I’m happy to participate.”
“One more thing, is it possible to get a copy of the URI-Rs that you guys detected so that I can feed them into an archive of my choice?...”

CC BY SA

Los Alamos NATIONAL LABORATORY

An Institutional Perspective to Rescue Scholarly Orphans

Martin Klein
Los Alamos National Laboratory
@mart1nkle1n
https://orcid.org/0000-0003-0130-2097

The Scholarly Orphans project
is funded by the Andrew W. Mellon Foundation

CC BY SA

Los Alamos NATIONAL LABORATORY

Scholarly Orphans – Pipeline

- 10,187 unique artifacts tracked, captured, and archived since 08/01/2018
- 41MB event database
- 61GB of WARC files
- 2.3GB of web archive index

CC BY SA

Los Alamos NATIONAL LABORATORY

Scholarly Orphans – Pipeline

- Capture process, post tracking
  - Within 9 minutes 50% of artifacts captured
  - Within 1 hour 21 minutes 75% of artifacts captured

- Archiver process, post capture
  - Within 10 minutes 50% of artifacts archived
  - Within 57 minutes 75% of artifacts archived

CC BY SA

Los Alamos NATIONAL LABORATORY